LA PRODUCTIVIDAD CIENTÍFICA DE LOS AUTORES

UN MODELO DE APLICACIÓN DE LA LEY DE LOTKA POR EL MÉTODO DEL PODER INVERSO GENERALIZADO

[THE SCIENTIFIC PRODUCTIVITY OF AUTHORS: AN APPLICATION MODEL OF LOTKA´S LAW BY THE GENERALIZED INVERSE POWER METHOD]

RUBÉN URBIZAGÁSTEGUI ALVARADO

Resumen: Describe la naturaleza de la distribución del poder inverso generalizado por el método de los mínimos cuadrados de la regresión lineal. Se describe paso a paso la forma de aplicación del modelo a los datos estudiados por Oliveira sobre la literatura de jaca. El modelo del poder inverso generalizado por el método de los mínimos cuadrados y la prueba Kolmogorov-Smirnov fueron usados para evaluar el ajuste de los datos observados y esperados. Con n = -2.14, C = 0.648515 y al 0.01 nivel de significación, se verificó que el valor crítico fue de 0.202177 con un desvío máximo de 0.151285, por lo tanto, esta literatura se ajusta al modelo de Lotka.

Palabras clave: Ley de Lotka; Distribución del poder inverso generalizado; Productividad de autores; Bibliometría; Jaca; Infometría; Cienciometría.

Abstract: The nature of the generalized inverse power distribution by the minimum least squared method is described. A step-by-step application process of this distribution model is offered. To illustrate the application process data collected and studied by Oliveira on Jackfruit literature was used. The generalized inverse power distribution by the minimum least squared method and the Kolmogorov-Smirnov test was used to evaluate the fit of the observed and expected data. With n = - 2.14 and C = 0.648715 at 0,01 level of significance, a critical value of 0.202177 with a maximum deviation of 0.151285, was observed. Therefore, the literature fits Lotka's model.

University of California, Riverside, University Libraries, P.O. Box 5900. Riverside, California 92517-5900. Estados Unidos de Norteamérica. Correo electrónico: ruben@ucr.edu
Artículo recibido: 10-02-05. Aceptado: 09-05-05
INFORMACIÓN, CULTURA Y SOCIEDAD. No. 12 (2004) p. 51-73

Keywords: Lotka's law; Generalized inverse power distribution; Author's productivity; Bibliometrics; Jackfruit; Infometrics; Scientometrics.

Introducción

La ciencia existe también como una práctica social colectiva y no solamente como una práctica individual. Ella es una realidad social objetivada en revistas, sociedades de especialistas, cátedras universitarias, bibliotecas y presupuestos específicamente dedicados a apoyar la ejecución de proyectos de investigación científica. Un problema que emerge constantemente en el mundo académico con relación a la práctica científica, es la productividad de sus participantes en la forma de publicaciones, tanto en sus aspectos cuantitativos como en los cualitativos. Esta tendencia ha dado lugar a reflexiones intelectuales sobre lo que ahora es considerado como la obligación de publicar y la existencia de un grupo de significantes contribuyentes en cualquier campo del conocimiento. Por lo tanto, se podría preguntar si la contribución de los grandes productores es de menor, igual, o mayor calidad que la contribución de los menores productores. Cattell (1910) afirmaba que no se sabía si el progreso de la ciencia se debía al gran número de trabajadores comunes o a la genialidad de unos pocos. Dennis (1955) analizando la productividad de los investigadores en lingüística, parálisis infantil, gerontología, geología y química, encontró que el 10% de los autores más prolíficos eran responsables por aproximadamente la mitad de todas las publicaciones en esas áreas, mientras que el 60% de los pequeños productores eran responsables por apenas 15% del total de las publicaciones. El propio Dennis (1954) estudiando el campo de la psicología encontró que los autores más productivos, eran también aquellos cuyos trabajos eran los más citados en otros trabajos académicos del campo de la psicología, por lo tanto, habría una asociación positiva entre la cantidad de trabajos publicados y el reconocimiento profesional en la forma de "visibilidad" académica. En consecuencia, se podría afirmar que, entre los autores, existen diferencias predeterminadas entre las habilidades y motivaciones para realizar trabajos creativos. Debido a diversos factores económicos y sociales, los autores más productivos tienen la tendencia a ser aún más productivos en el futuro, mientras que los autores menos productivos muestran una tendencia a declinar en productividad.

Los estudios sobre la productividad de los autores no son privativos de la Bibliotecología y la Ciencia de la Información sino que también son hechas por psicólogos y sociólogos pero en distintas direcciones. Mientras que los psicólogos están más interesados en explorar el mundo de la creatividad, los factores cognitivos que hacen posible la existencia de los "genios" y la "inteligencia", los

sociólogos apuntan a las condiciones sociales que hacen posible la producción estratificada y desigual en la ciencia. En cambio, los bibliotecólogos están más interesados en las "publicaciones" (tesis, libros, artículos, etc.) como un producto acabado y objetivado de la práctica científica. Eso hace imposible hablar de la Ley de Lotka sin hablar de la bibliometría como una disciplina en busca de consolidación, autonomía y legitimación como campo científico.

Sin duda, el principal objetivo de cualquier ciencia es establecer principios generales que puedan explicar y predecir el comportamiento de los fenómenos estudiados, por eso la necesidad de elaborar leyes de desarrollo y teorías. En consecuencia, la práctica científica requiere la obtención de datos o evidencias empíricas por un lado y la elaboración de teorías por otro lado. Por lo tanto, cualquier práctica científica debe identificar y delimitar su objeto de estudio y sus problemas de investigación; debe también descubrir leyes empíricas que expresen cierto tipo de relación entre los fenómenos observados, así como formular una estructura sistemática que contengan las leyes empíricas y que las expliquen de manera racional.

Siguiendo los presupuestos epistemológicos establecidos anteriormente, los estudios sobre la productividad de autores hacen su aparición en la década de los 20s. Dresden (1922) fue la primera persona en prestar atención al fenómeno de la producción bibliográfica. Estudió la producción de artículos de autores ligados a la Sociedad Americana de Matemáticas, Sección de Chicago, de 1897 a 1922. Cuatro años mas tarde Lotka (1926) intentando determinar la parte con que los autores contribuyen al progreso de la ciencia, contó el número de nombres que aparecían en el *Chemical Abstract, 1907-1916* y el *Auerbach Geschichtstafeln der Physic,* hasta 1900. Después trazó la frecuencia de personas que efectuaban 1, 2, 3, etc. contribuciones, frente al número de 1, 2, 3, etc. contribuyentes con ambas variables en escala logarítmica. Encontró que los puntos estaban estrechamente esparcidos sobre una línea recta teniendo una inclinación de aproximadamente igual a 2, por consiguiente, concluyó que la fórmula general para la relación entre la frecuencia *y* de las personas que efectuaban *x* contribuciones era $x^n y = const$, y la proporción de los autores que contribuían con un único ítem era de más o menos el 60 por ciento.

Doce años después Dufrenoy (1938) examinó el comportamiento de publicación de los biólogos y sugirió que estos datos reflejaban la ley de Lotka solo para los pequeños valores de las *n* contribuciones. Hersh (1942) usando una exhaustiva bibliografía publicada en 1939 por H. J. Muller sobre la genética de Drosophila analizó el número de artículos publicados cada año desde 1910 hasta 1938 según el número de autores. Cuando el número de artículos publicados fueron trazados en una escala semilogarítmica frente al tiempo transcurrido, encontró que los puntos se agrupaban dentro de una estrecha banda con bordes rectos y paralelos, es decir, los datos se ajustaban a la simple relación exponencial llamada de ley del interés compuesto. Cuando el logaritmo del

número de autores se trazaron frente al logaritmo del número de artículos acreditados a los autores, los diez primeros puntos (que agrupaban a cerca del 90% de los autores) se alineaban en línea recta con una inclinación negativa que se asemejaba a la ley de Lotka. En 1944 Williams (1944) discutía los datos de Dufrenoy (1938) y adjuntó otros dos datos sobre biólogos. Sus resultados fueron similares a los de Dufrenoy con – 0.40 y – 0.31 como cálculos de las inclinaciones respectivamente. Leavens (1953) evaluó los trabajos de econometras usando *Guide to Econometrica* de diciembre de 1952. Esta guía contenía información pertinente a los artículos publicados en los primeros veinte volúmenes de *Econométrica* (1933-1952). Sus datos produjeron resultados similares a la curva de Pareto sobre ingresos familiares, que tiene una inclinación recta. Esta información se relacionaba a la generalización de Zipf, y puede encontrarse en diferentes fenómenos económicos y sociales. Él discutió el hecho de que el gráfico no se ajustaba a una línea recta en toda su extensión. Sin embargo, trazando la línea a mano, casi todos los puntos, con excepción de los tres últimos, se ajustaban estrechamente a la teoría de Pareto (con una inclinación de – 1.5 aproximadamente). Aunque Leavens se dio cuenta que los artículos no producían lo que de ellos se esperaba, explicó que *Econométrica* era una entre muchas otras fuentes de publicación disponibles para los 721 autores contribuyentes.

El modelo del Lotka ha sido probado en muchas áreas que incluyen bases de datos de patentes (Oppenheim, 1986), aceites lubricantes (López Calafi, Salvador y Guardia, 1985), educación superior (Budd, 1988), música popular (Cook, 1989), finanzas (Chung and Cox, 1990), economía (Cox and Chung, 1991), contabilidad (Chung, Pak and Cox, 1992), industria musical (Cox, Felton and Chung, 1995), psiquiatría (López-Muñoz y Rubio Valladolid, 1995), glándula pineal y melatonina (López-Muñoz; et al., 1996), bibliotecología y ciencia de la información española (Jiménez Contreras y Moya-Anegón, 1997), genética (Gupta, Kumar y Rousseau, 1998), bibliometría (Urbizagástegui Alvarado, 1999), geología (Urbizagástegui Alvarado y Cortés, 2002) y ha sido explorado en la propia literatura de Lotka (Urbizagástegui Alvarado, 2002). Pero los datos de estas investigaciones varían mucho yendo desde datos tomados de bibliografías exhaustivas como las de entomología (Gupta, 1987), investigación sobre la papa (Gupta; et al., 1996) hasta datos tomados de un grupo de revistas (Nath and Jackson, 1991) y hasta de una única revista. Algunos estudios usaron datos sobre la historia de un tema, mientras que otros consideraron solo datos sobre pocos años. Debe recalcarse que la ley de Lotka ha sido probada frente a muchas recopilaciones de datos, y el ajuste no siempre ha sido bueno. Este punto no ha sido recalcado lo suficiente en la bibliografía publicada, aunque las revisiones del estado del arte de Vlachy (1980) y Potter (1981) han ayudado a reorientar el interés de los investigadores para evaluar si otros tipos de distribuciones pueden proporcionar un mejor ajuste de los datos. Esto llevó a introducir

en las investigaciones modelos estadísticos usados en las ciencias naturales tales como la distribución hiperbólica, la distribución logarítmica normal, la distribución de Yule, la distribución binomial, la distribución negativa binomial, la serie geométrica, la serie logarítmica, la distribución de Weinbull, la distribución de Poisson, la distribución truncada de Poisson y finalmente la distribución inversa generalizada de Gauss-Poisson.

A pesar de que esos modelos estadísticos están disponibles en la literatura publicada y son accesibles a través de los textos de enseñanza de estadística, no han sido suficientemente explorados en la bibliometría latinoamericana. Tal vez eso se deba a la falta de familiaridad con los modelos estadísticos o al desconocimiento de los instrumentos matemáticos para probar los datos. Es sabido que desentrañar complejos modelos estadísticos no es el pasatiempo favorito de ningún bibliotecario. Por esa razón, el objetivo de este trabajo es proporcionar una guía didáctica para analizar y probar la ley de Lotka sobre la productividad científica de los autores usando el modelo de los mínimos cuadrados. Consideramos que esta guía es necesaria para apoyar la comprensión, adopción, aplicación y difusión de este modelo.

La distribución del poder inverso generalizado

La ley de Lotka es una distribución de probabilidades discretas que describe la productividad de autores. Originalmente propuesta por Lotka (1926) como un modelo del cuadrado inverso, ahora es conocido como la ley de Lotka, una forma más general llamado de poder inverso generalizado y que es expresado en la forma de:

$$y_x = C\, x^{-n}, \qquad x = 1, 2, \ldots, x_{max} \qquad (1)$$

donde,

y_x es la probabilidad de que un autor haga $\boldsymbol{x}$ contribuciones sobre un asunto

C y n son los dos parámetros que deben ser estimados de los datos observados.

Los principales elementos envueltos en el "ajuste" del modelo del poder inverso generalizado son los siguientes:

1. *Medición y tabulación*

El número de autores $\boldsymbol{y_x}$ contribuyendo $\boldsymbol{x}$ artículos en un asunto determinado deben ser organizados en una tabla de frecuencias decrecientes de $\boldsymbol{N}$ pares $\boldsymbol{x, y}$.

Las medidas de productividad de los autores deben tomar en consideración a todos los autores incluyendo a los colaboradores. Aparentemente, Lotka solamente contó los primeros autores porque múltiples autores eran menos comunes en ese tiempo, y probablemente porque de esa forma el conteo era más fácil. En un mundo post-moderno y de capitalismo avanzado, la investigación se caracteriza por una extensiva colaboración que se refleja en los múltiples autores de un único trabajo. Si estamos interesados en la distribución de la productividad de esos autores, las medidas que no tomen en consideración esas colaboraciones y que no sean sensibles a este fenómeno son inválidas.

2. *Modelo adoptado*

El modelo adoptado es el poder inverso generalizado en la forma de:

$$\gamma_x = C\left(\frac{1}{x^n}\right) \qquad (2)$$

que es la forma que adopta la ecuación (1) cuando se elimina de ésta ecuación el signo negativo del exponente *n*.

3. *Estimación del parámetro n*

Es calculado usando el método del mínimo cuadrado linear y expresado como:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

donde

N = número de pares de datos observados
X = logaritmo de base 10 de x
Y = logaritmo de base 10 de y

4. *Estimación del parámetro C*

Para estimar C se usa la función inversa Zeta de Riemann. Para esa estimación Pao (1985, 1986) proporciona una fórmula de aproximación exacta expresada como:

$$C = \frac{1}{\sum_{x=1}^{P-1} \frac{1}{x^n} + \frac{1}{(n-1)\ P^{n-1}} + \frac{1}{2\ P^n} + \frac{n}{24\ (P-1)^{n+1}}}$$

donde,

P = es el número de pares de datos *xy* observados.

5. Prueba del ajuste

Para investigar el ajuste de la distribución se usa la prueba Kolmogorov-Smirnov (K-S), que es aplicado al conjunto de valores observados y esperados a un 0.01 nivel de significación. Esta prueba compara la distribución de las frecuencias observadas con la distribución de las frecuencias calculadas o teóricas, usando la función acumulada de ambas distribuciones. Una de sus ventajas es que trabaja muy bien con pequeñas muestras, no pierde información con la agrupación de los datos de la distribución en clases como lo hace la prueba de chi-cuadrado y es más poderosa que esa prueba de chi-cuadrado.

Aplicación de la distribución del poder inverso generalizado por el método de los mínimos cuadrados

Vamos a proporcionar un ejemplo de aplicación de la Ley de Lotka por el método del poder inverso generalizado usando el modelo de los mínimos cuadrados. Los datos proceden de un estudio realizado por Oliveira (1983), sobre la literatura de Jaca. Para eso, se recomienda seguir las siguientes etapas.

1. Recolección de los datos y distribución de las frecuencias observadas

Se sugiere tomar una bibliografía existente (o elaborar una), cuya cobertura sea extensa (cuanto más extensa, en tiempo, mejor). Se sugiere una cobertura de diez años o más. Haga el conteo del número de contribuciones hechas por cada autor incluyendo los coautores. Ordene esos datos en una tabla de ocho columnas. En las dos primeras columnas indique los valores de *x* (el número de las contribuciones, ex. 1, 2, 3, 4 ... etc.) en orden creciente, y los valores de *y* (el número de autores que hayan hecho x contribuciones) ordenados de forma decreciente; por lo tanto, el primer x (una contribución) deberá contener el mayor número de autores con una contribución. Aquí la primera previsión a ser tomada debe ser completar con cero las frecuencias de los autores cuyo número de contribuciones no fueron observados.

En la opinión de Loughner (1992), de lo que se trata es de evitar que existan células vacías que más adelante puedan afectar la prueba Kolgomorov-Smirnov. Por ejemplo, en el caso de los datos de Oliveira (1983), no se observó ningún autor con 5 colaboraciones y por eso esa quinta colaboración fue completada con cero número de autores. En la tercera columna indique los valores de *xy* (multiplicación de x por y). Esta columna indica el número de artículos producidos por los autores. En la cuarta columna indique Sxy (la suma acumulada de la multiplicación de *x* por *y*). En la quinta columna indique el porcentaje de y, es decir, el porcentaje de los autores. En la sexta columna indique la S%*y* (el valor acumulado del porcentaje de y de la quinta columna). En la séptima columna el porcentaje del total de los artículos (% de *xy* de la tercera columna). En la octava columna indique los valores acumulados de la séptima columna.

Esta primera tabla sirve solo para una observación general de los valores obtenidos. Calcule con ella la productividad media por autor, describa el porcentaje de autores con una única contribución, es decir, algún aspecto interesante que resaltar en la distribución. Por ejemplo, en los datos de Oliveira (1983), mostrados en la *Tabla 1*, existe una concentración del 80% de autores contribuyendo con un solo artículo. Este porcentaje es 20% más alto que los 60% pronosticados por Lotka. En el lado opuesto, solo 5% de los autores produjeron 4 o más artículos. La producción media es de 1.4 con una varianza de 1.02 artículos por autor.

1	2	3	4	5	6	7	8
No. de contribuciones por autor	*No. de autores*	*total de artículos*		*% de autores*		*% de artículos*	
x	**y**	**x.y**	**Σ x.y**	**% y**	**Σ % y**	**% xy**	**Σ % xy**
1	52	52	52	80.0	80.0	58.8	58.8
2	8	16	68	12.5	92.5	17.8	76.6
3	2	6	74	3.0	95.5	6.7	83.3
4	1	4	78	1.5	97.0	4.4	87.7
5	0	0	78	0.0	97.0	0.0	87.7
6	2	12	90	3.0	100.0	13.3	100.0
	65	**90**		**100.0**		**100.0**	

Tabla 1: Frecuencia observada de contribuciones por autor

2. Tabla de los mínimos cuadrados

Elabore otra tabla conteniendo de nuevo seis columnas. Las columnas 1 y 2 contienen los mismos valores de *x* e *y* conforme descrito en el punto 1. La columna 3 contiene el logaritmo de los valores de *x* de la columna 1. La columna cuatro contiene los valores del logaritmo de *y* de la columna 2. La columna

cinco contiene los valores de la multiplicación de los logaritmos de *x* por *y* (columna 3 por columna 4). La columna seis contiene los valores del logaritmo de *x* elevados al cuadrado (columna 3 elevados al cuadrado). La *Tabla 2* incluye un ejemplo usando los datos de la productividad de los autores que produjeron artículos sobre jaca estudiados por Oliveira (1983).

1	2	3	4	5	6
No. de contribuciones por autor **x**	*No. de autores* **y**	**log x.**	**log y**	**log x (log y)**	**(log x)^2**
1	52	0.00000	1.71600	0.00000	0.00000
2	8	0.30103	0.90309	0.27186	0.09062
3	2	0.47712	0.30103	0.14363	0.22764
4	1	0.60206	0.00000	0.00000	0.36248
5	0	0.69897	0.00000	0.00000	0.48856
6	2	0.77815	0.30103	0.23425	0.60556
Total	**65**	**2.85733**	**3.22115**	**0.64974**	**1.77486**

Tabla 2: Distribución de los mínimos cuadrados de los datos observados

3. Trazado del gráfico

Sobre un papel en escala normal trace el logaritmo de *y* frente al logaritmo de *x*. Algunos autores recomiendan inspeccionar visualmente el trazado para determinar el punto de corte de la línea recta. Este es el llamado método del "ojímetro". Recomiendan también cortar los valores de los grandes productores que se desvían demasiado de la línea recta. Otros autores recomiendan no ejecutar ningún corte y trabajar con todos los autores (incluidos los coautores) presentes en la distribución. Como podemos ver en la *Figura 1*, con los datos completos (sin corte alguno) la regresión linear entre los artículos producidos y los autores productores de artículos, es del 85.36%. Sin embargo, lo más recomendable es trabajar con todos los autores, es decir, sin corte alguno. Especialmente porque el cálculo del coeficiente de correlación de Pearson al cuadrado r^2 puede ser hecho fácilmente con paquetes estadísticos automatizados tales como Excel, Notebook, Minitab, SPSS, SAS, Mathematica y otros disponibles en el mercado.

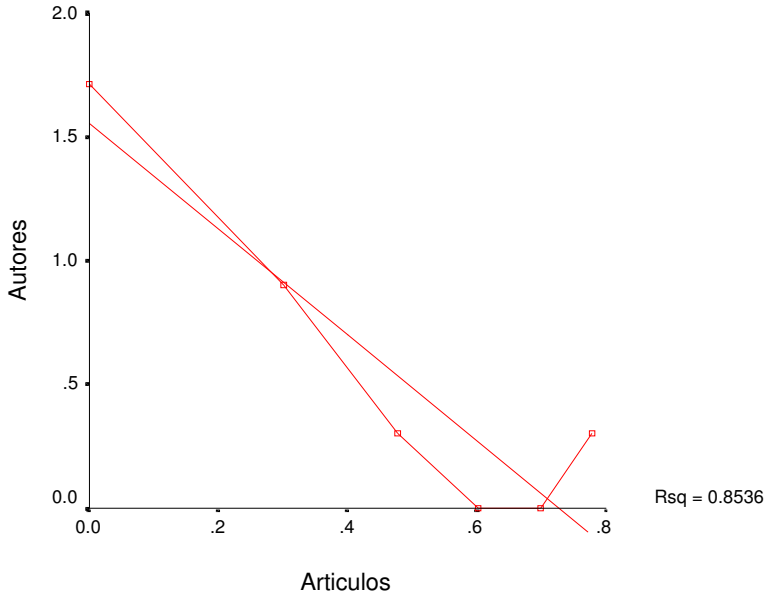

Figura 1: Recta de regresión del coeficiente de correlación de Pearson al cuadrado

Ahora, si cortamos el último autor de mayor producción la correlación entre autores y artículos producidos aumenta hasta alcanzar el 97.61 %. Precisamente, Pao (1986) recomienda calcular la regresión linear o el coeficiente de determinación r^2 removiendo paulatinamente cada vez uno de los datos de los autores más productivos hasta alcanzar el mayor porcentaje de r^2. Por ejemplo, si cortamos los datos de los 2 autores que produjeron 6 artículos, el r^2 aumenta al 97.61%. Este coeficiente de determinación permite establecer la cantidad de variación en la variable g que es explicada por la ecuación de regresión producida por la variable *x*. Como se muestra en la *Figura 2*, el objetivo es encontrar la "mejor" línea de regresión para los datos

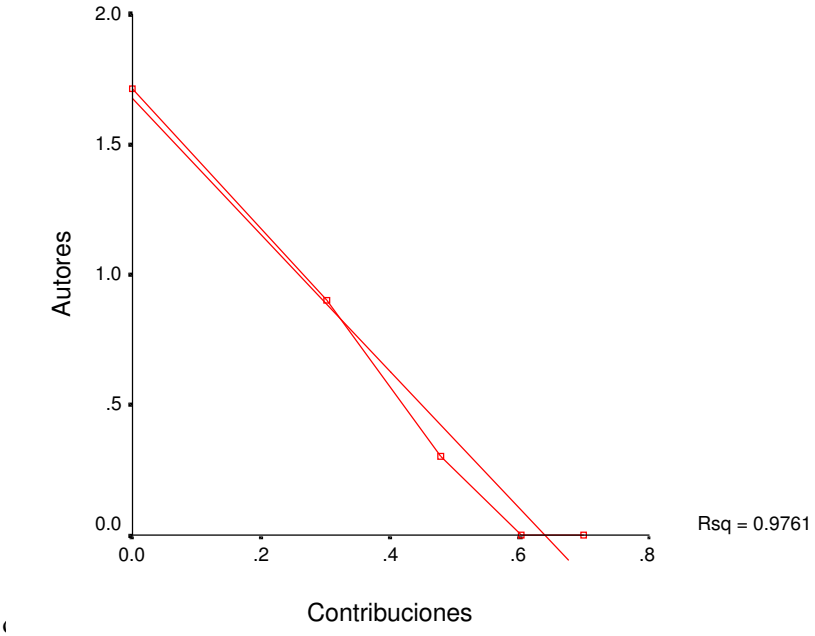

Figura 2: Recta de regresión truncada del coeficiente de correlación de Pearson al cuadrado

4. Cálculo del valor de n

Utilizando los valores mostrados en la *Tabla 2* de los mínimos cuadrados elaborado en el punto 2, hacer el cálculo del "mejor" valor de la caída de *n*. Esta caída es el valor del exponente *n* de la Ley de Lotka. Para calcular el valor de ***n*** use la siguiente ecuación:

$$n = \frac{N\sum XY - \sum X \sum Y}{N\sum X^2 - \left(\sum X\right)^2}$$

donde

N = número de pares de datos observados
X = log x (base 10)
Y = log y (base 10)

Para ilustrar el cálculo del parámetro *n*, vamos usar los datos de Oliveira (1983), tomando los valores ya calculados en la *Tabla 2* de la distribución de los mínimos cuadrados. En este caso vamos a usar todos los datos sin corte alguno, que es lo más recomendable.

$$n = \frac{6(0.64974) - (2.85733)(3.22115)}{6(1.77482) - (2.85733)^2}$$

$$n = \frac{3.89844 - 9.20389}{10.64892 - 8.16433}$$

$$n = \frac{-5.30545}{2.48459}$$

n = - 2.135

redondeando,

n = - 2.14

5. Cálculo de C

C representa el porcentaje teórico de los autores colaborando con un único artículo o trabajo en la distribución de la productividad de autores. Haga el cálculo de C, substituyendo el valor del parámetro *n* obtenido en la etapa 4, y usando P = 6 (porque son seis pares de datos), en la siguiente ecuación:

$$C = \frac{1}{\sum_{x=1}^{P-1} \frac{1}{x^n} + \frac{1}{(n-1)\,P^{n-1}} + \frac{1}{2\,P^n} + \frac{n}{24\,(P-1)^{n+1}}}$$

donde,

x = es el número de 1, 2, 3, …n contribuciones por autor.
n = es el valor del parámetro *b* estimado en el punto 4. En este caso, *n* es igual a - 2.14
P = es el número de pares de datos observados. Como se muestran en la *Tabla 2*, en este caso, el valor de P es igual a 6.

El valor de P variará dependiendo del número de pares *x e y* observados. Si los pares de datos observados fuesen 10, 15, 16, etc. el valor de P también será igual a 10, 15 o 16. Por ejemplo, continuando con los datos de Oliveira (1983):

$$C = \frac{1}{\sum_{x=1}^{P-1} \frac{1}{x^n} + \frac{1}{(2.14-1)\,(6)^{2.14-1}} + \frac{1}{2\,(6)^{2.14}} + \frac{2.14}{24\,(6-1)^{2.14+1}}}$$

El denominador de esta ecuación esta compuesta por la suma de 4 ecuaciones diferentes. Así que vamos a calcularlos una por una:

a) Primera ecuación

La ecuación $\sum_{x=1}^{P-1} \frac{1}{x^n}$ que es la primera parte del denominador, indica una sumatoria de los *x* que van desde *x* = 1 hasta P - 1, y si P = 6, entonces, P – 1 = 6 – 1 = 5. Después, esta ecuación es resuelta de la manera siguiente:

$$\sum_{x=1}^{P-1} \frac{1}{x^n} = \frac{1}{1^{2.14}} + \frac{1}{2^{2.14}} + \frac{1}{3^{2.14}} + \frac{1}{4^{2.14}} + \frac{1}{5^{2.14}}$$

$$\sum_{x=1}^{P-1} \frac{1}{x^n} = \frac{1}{1} + \frac{1}{4.40762} + \frac{1}{10.4964} + \frac{1}{19.4271} + \frac{1}{31.3181}$$

$$\sum_{x=1}^{P-1} \frac{1}{x^n} = 1.40556$$

b) Segunda ecuación

$$\frac{1}{(2.14-1)\,(6)^{2.14-1}} = \frac{1}{1.14\,(7.71068)} = \frac{1}{8.79018} = 0.113763$$

c) Tercera ecuación

$$\frac{1}{2\,(6)^{2.14}} = \frac{1}{92.52811} = 0.021615$$

c) Cuarta ecuación

$$\frac{2.14}{24\,(6-1)^{2.14+1}} = \frac{2.14}{24\,(5)^{3.14}} = \frac{2.14}{24\,(156.591)} = \frac{2.14}{3758.18} = 0.000569425$$

e) Por último, sumando los valores:

$$1.4056 + 0.113763 + 0.021615 + 0.000569425 = 1.54151$$

f) Finalmente,

$$C = \frac{1}{1.54161} = 0.648715$$

6. Cálculo de los valores esperados o teóricos

Con los valores de los parámetros n = - 2.14 y C = 0.648715 ya conocidos, calcular las frecuencias esperadas o teóricas usando la ecuación (1)

$$y_x = C\,x^{-n}, \qquad x = 1, 2, \ldots, x_{max} \qquad (1)$$

Esta ecuación es una simple multiplicación de la constante C por el número de contribuciones en artículos *x* elevados al valor del exponente negativo de *n*. De modo que para eliminar el signo negativo de *n* y resolverlo se procede de la siguiente manera:

$$y_x = C\left(\frac{1}{x^n}\right)$$

Después, usando esta ecuación se procede a calcular los valores esperados o teóricos de la distribución de frecuencias.

a) Para $x = 1$ (el número de autores que produjeron 1 artículo)

$$y_1 = 0.648715 \times \frac{1}{1^{2.14}} = 0.648715 \times 1 = 0.648715$$

Ahora, $0.648715 \times 65 = 42.166475 \pm 42.2$

b) Para $x = 2$ (el número de autores que produjeron 2 artículos)

$$y_2 = 0.648715 \times \frac{1}{2^{2.14}} = 0.648715 \times \frac{1}{4.40762} = 0.14718$$

Ahora, $0.14718 \times 65 = 9.5667 \pm 9.6$

c) Para $x = 3$ (el número de autores que produjeron 3 artículos)

$$y_3 = 0.648715 \times \frac{1}{3^{2.14}} = 0.648715 \times \frac{1}{10.4964} = 0.0618036$$

Ahora, $0.0618036 \times 65 = 4.01723 \pm 4.0$

d) Para $x = 4$ (el número de autores que produjeron 4 artículos)

$$y_4 = 0.648715 \times \frac{1}{4^{2.14}} = 0.648715 \times \frac{1}{19.4271} = 0.0333922$$

Ahora, $0.0333922 \text{ ´ } 65 = 2.17049 \pm 2.2$

e) Para $x = 5$ (el número de autores que produjeron 5 artículos)

$$y_5 = 0.648715 \times \frac{1}{5^{2.14}} = 0.648715 \times \frac{1}{31.3181} = 0.0207137$$

Ahora, $0.0207137 \times 65 = 1.34639 \pm 1.3$

f) Para $x = 6$ (el número de autores que produjeron 6 artículos)

$$y_6 = 0.648715 \times \frac{1}{6^{2.14}} = 0.648715 \times \frac{1}{46.2641} = 0.014022$$

Ahora, $0.014022 \times 65 = 0.91143 \pm 0.9$

7. Establecer las hipótesis

Lo que se va a probar es si la distribución del poder inverso generalizado obtenida experimentalmente por el método de los mínimos cuadrados es homogénea o no. Es decir, la probabilidad de que un elemento incluido en la muestra, es la misma (igualmente probable) para todos los elementos en esa misma situación. Por lo tanto, establecemos las hipótesis de la siguiente manera:

H_o = la distribución representa los conteos de $x = 1, 2, 3$ ..... artículos
Ha ≠ la distribución no representa los conteos de $x = 1, 2, 3$ ..... artículos

8. Especificar la región de rechazo de las hipótesis al nivel de significación de $\alpha = .01$

Usando el nivel de significación de a = .01 en la tabla de los valores críticos de la prueba K-S de cualquier texto estadístico (*Tabla 1 en el Anexo*) encontrar la región de rechazo. Mirando en esa tabla en la columna de los $n = 65$ (n es el tamaño de la muestra) y en la columna del nivel de significación de a = .01, encontramos que para una muestra de $n = 65$ autores, el valor crítico de la desviación máxima debe ser calculada usando la siguiente fórmula:

$$\frac{1.63}{\sqrt{n}}$$

Esto significa dividir 1.63 entre la raíz cuadrada de la población total de autores en estudio y listados en la tabla de la distribución, esto es:

$$\frac{1.63}{\sqrt{65}} = \frac{1.63}{8.06226} = 0.202177$$

Por lo tanto, el valor crítico de la prueba K-S es igual a 0.202177.

9. Prueba de ajuste Kolmogorov-Smirnov de la distribución teórica de la productividad de autores

La prueba de ajuste Kolmogorov-Smirnov (K-S) es un simple método no-paramétrico para probar si hay diferencias significativas entre las frecuencias observadas y las frecuencias teóricas o calculadas de una distribución. Es una medida de la bondad del ajuste de una distribución de frecuencias similar al c^2 (chi-cuadrado). Sin embargo, esta prueba K-S, es más poderosa que el c^2 (chi-cuadrado), más fácil de usar y no necesita que los datos estén agrupados en frecuencias inferiores a 5 como lo exige la prueba chi-cuadrada. Es particularmente útil para juzgar cuán próximas están las frecuencias observadas de las frecuencias calculadas o esperadas.

Con los valores de *n* y *c* obtenidos en las etapas 3 y 4, construir otra tabla con siete columnas. La columna 1 y 2 contienen los valores de *x* e *y* respectivamente. La columna 3, contiene el porcentaje de autores haciendo 1, 2, 3, 4, y así por delante, contribuciones en los datos observados. La columna 4, contiene los valores acumulados de la columna 3. La columna 5, contiene los valores teóricos computados usando *n* y *c* con la formula C ($1/x^n$). La columna 6, contiene los valores acumulados de la columna 5. La columna 7 contiene la diferencia máxima absoluta (Dmax) entre los pares de valores de la columna 4 y la columna 6.

En la columna 7 identifique la desviación máxima *(Dmax)*. Este es el máximo valor absoluto encontrado en las diferencias de los valores de las columnas 4 y 6. Valor absoluto significa tomar el valor calculado sin considerar los signos negativos de la desviación máxima calculada. En el caso del trabajo de Oliveira (1983), la Dmax es igual a 0.151285, resaltado en negritas en la *Tabla 3* siguiente.

1	2	3	4	5	6	7
x	*y*	$y_x/\Sigma y_x$	$\Sigma(y_x/\Sigma y_x)$	$C\ (1/x^n)$	$\Sigma\ [C\ (1/x^n)]$	*Dmax*
1	52	.800000	0.800000	0.648715	0.648715	**0.151285**
2	8	.123077	0.923077	0.147180	0.795895	0.127182
3	2	.030769	0.953846	0.061804	0.857699	0.096147
4	1	.015385	0.969231	0.033392	0.891091	0.078140
5	0	.000000	0.969231	0.020714	0.911805	0.057426
6	2	.030769	1.000000	0.014022	0.925827	0.074173

Tabla 3: Prueba de ajuste Kolmogorov-Smirnov de la distribución de los autores productores de literatura sobre Jaca

Después compare el tamaño del valor crítico de la prueba K-S encontrado, con el valor de la desviación máxima (Dmax) calculada. De modo que ahora solo existen dos alternativas: mayor o menor. Es el valor crítico de 0.202177 mayor o menor que el valor de la desviación máxima (Dmax) 0.151285? En este caso, el valor crítico 0.202177 es mayor que la desviación máxima (Dmax) 0.151285. Por lo tanto, la región de rechazo de las hipótesis es la parte puntillada y obscurecida de la siguiente *Figura 3*. La región de aceptación es la parte a la izquierda de esta figura.

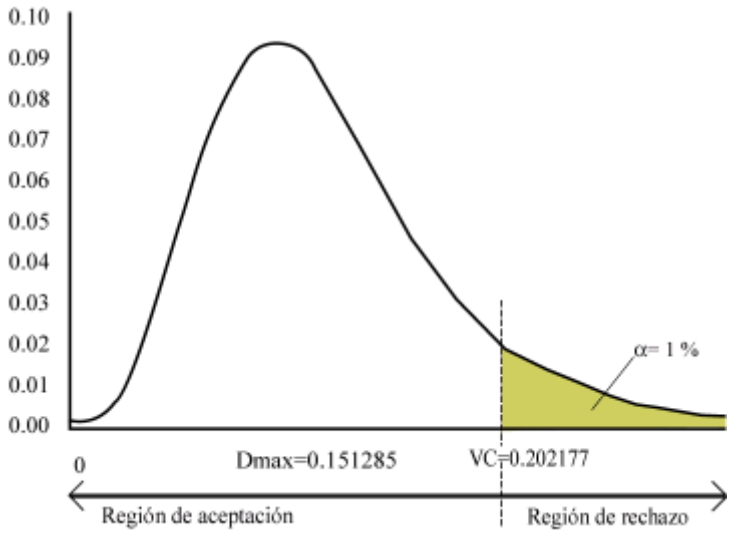

Figura 3: Gráfico de la región de aceptación y rechazo de las hipótesis

10. Interpretación del ajuste de la distribución a la Ley de Lotka

Ya que se tienen solamente dos alternativas, solo hay dos formas de interpretación: la distribución se ajusta a la Ley de Lotka o la distribución no se ajusta a la Ley de Lotka.

a) Si la desviación máxima *(Dmax)* es mayor que el valor crítico (i.e. el valor crítico es menor que la (*Dmax*), rechace la hipótesis nula de homogeneidad de la distribución de frecuencias de los productores de la literatura de Jaca, es decir, rechazar la hipótesis de que esta distribución se ajusta a la Ley de Lotka al 0.01 nivel de significación.

b) Si la desviación máxima *(Dmax)* es menor que el valor crítico (i.e. el valor crítico es mayor que la desviación máxima *(Dmax*), acepte la hipótesis nula de homogeneidad de la distribución de frecuencias de los productores de la literatura de Jaca, es decir, aceptar la hipótesis de que esta distribución se ajusta a la Ley de Lotka al 0.01 nivel de significación.

Como en este caso, el valor crítico 0.202177 es mayor que la Dmax 0.151285, se acepta la hipótesis nula y se concluye que esta distribución se ajusta a la Ley de Lotka a un 0.01 nivel de significación.

11. Comparación de los valores observados y los valores calculados

Elabore una tabla especial para comparar los valores observados y calculados. Observar en esta tabla cuán próximos o alejados están ambos valores. Observar también que los totales son los mismos, casi los mismos, o divergentes para ambas frecuencias. Los resultados para los dados de la literatura de Jaca estudiados por Oliveira (1983) y replicados en este trabajo están mostrados en la *Tabla 4* siguiente:

Contribuciones por autor x	*Frecuencia observada*	*Frecuencia esperada*
1	52	42.2
2	8	9.6
3	2	4.0
4	1	2.2
5	0	1.3
6	2	0.9
Total	65	60.2

Tabla 4: Frecuencias observadas y esperadas

12. Gráfico de dispersión de los valores observados y esperados

Como se muestra en la *Figura 4* siguiente, la aproximación entre los valores observados y esperados de la distribución de la productividad de autores pueden ser mejor observadas en el trazado de la dispersión de ambos valores. Esta figura debe ser incluida en la redacción del informe final.

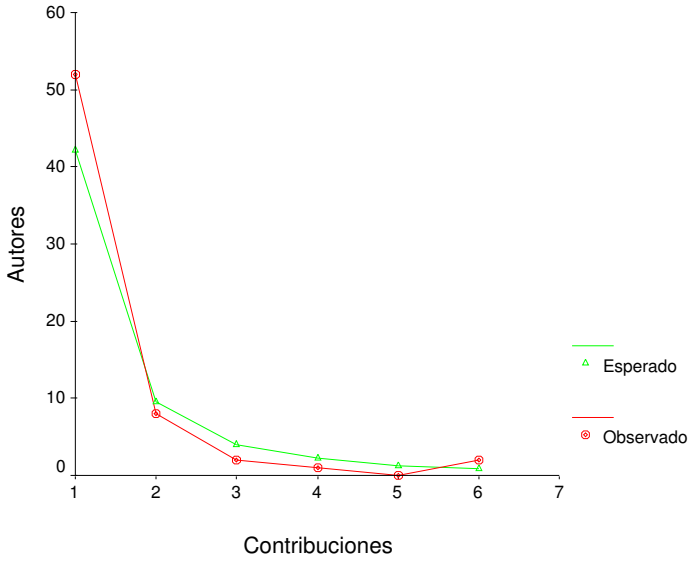

Figura 4: Dispersión de las frecuencias observadas y esperadas

Conclusiones

El modelo permitió identificar un alto porcentaje de 92.5% de pequeños productores responsables por 77% de la literatura publicada. Por su parte los medianos y grandes productores de la literatura de jaca corresponden apenas al 7.5% de los autores responsables por 23% de la literatura producida.

Fueron encontrados 65 autores que conjuntamente produjeron 90 artículos, siendo que 80% de ellos contribuyeron con un único artículo a la literatura estudiada. El modelo del poder inverso generalizado por el método de los mínimos cuadrados y la prueba Kolmogorov-Smirnov fueron usados para evaluar el ajuste de los dados observados y esperados. Con n = -2.14, C = 0.648515 y al 0.01 nivel de significación, se verificó que el valor crítico fue de 0.202177 con una desviación máxima de 0.151285, por lo tanto, esta literatura se ajusta muy bien al modelo de Lotka.

Estos resultados son diferentes a los informados por Oliveira (1983) quién afirmó que “la ley de Lotka no se aplica a la literatura de Jaca”. El autor encontró también que el valor del exponente n fue igual a 3.02 más próximo a lo determinado por Price que a la formulación original de Lotka. Tal vez sea

pertinente recordar que el autor no estimó el valor del exponente *n* de sus datos observados, sino que a priori estableció el valor de *n* como siendo igual a 2. Tampoco probó estadísticamente sus datos. Tal vez eso explica las divergencias encontradas con este trabajo.

Referencias bibliográficas

Budd, John M. 1988. A bibliometric analysis of higher education literature. En *Research in Higher Education.* Vol. 28, no. 2, 180-190.

Cattell, J. M. 1910. A further statistical study of American men of science. En *Science*. Vol. 32, 633-648; 672-688.

Chung, Kee H. and Raymond A. K. Cox. 1990. Patterns of productivity in the finance literature: a study of the bibliometric distributions. En *The Journal of Finance.* Vol. 45, no. 1, 301-309.

Chung, Kee H.; Hong S. Pak and Raymond A. K. Cox. 1992. Patterns of research output in the accounting literature: a study of the bibliometric distributions. En *Abacus.* Vol. 28, no. 2, 168-185.

Cook, Kevin L. 1989. Laws of scattering applied to popular music. En *Journal of the American Society for Information Science.* Vol. 40, no. 4, 277-283.

Cox, Raymond A. K. and Kee H. Chung. 1991. Patterns of research output and author concentration in the economics literature. En *The Review of Economics and Statistics*. Vol. 73, no. 4, 740-747.

Cox, Raymond A. K.; James M. Felton and Kee H. Chung. 1995. The concentration of commercial success in popular music: an analysis of the distribution of gold records. En *Journal of Cultural Economics.* Vol. 19, no. 4, 333-340.

Dennis, Wayne. 1954. Productivity among American psychologists. En *American Psychologist.* Vol. 9, no. 5, 191-194.

Dennis, Wayne. 1955. Variations in productivity among creative workers. En *The Scientific Monthly.* Vol. 80, no. 4, 277-278.

Dresden, A. 1922. A report on the scientific work of the Chicago Section, 1897-1922. En *Bulletin of The American Mathematical Society.* Vol. 28, 303-307.

Dufrenoy, Jean. 1938. The publishing behavior of biologists. En *Quarterly Review of Biology.* Vol. 13, 207-210.

Gupta, B. M.; Suresh Kumar and R. Rousseau. 1998. Applicability of selected probability distributions to the number of authors per article in theoretical population genetics. En *Scientometrics.* Vol. 42, no. 3, 325-334.

Gupta, B. M.; Suresh Kumar; Shaheen Syed and Karan Vir. Sing. 1996. Distribution of productivity among authors in potato research (1900-1980). En *Library Science with a Slant to Documentation.* Vol. 33, no. 3, 127-134.

Gupta, Davendra K. 1987. Lotka's law and productivity patterns of entomological research in Nigeria for the period, 1900-1973. En *Scientometrics.* Vol. 12, no. 1-2, 33-46.

Hersh, A. H. 1942. Drosophila and the course of research. En *Ohio Journal of Science.* Vol. 42, no. 5, 198-200.

Jiménez Contreras, Evaristo y Félix Moya-Anegón. 1997. Análisis de la autoría en revistas españolas de Biblioteconomía y Documentación, 1975-1995. En *Revista Española de Documentación Científica.* Vol. 20, no. 3, 252-266.

Leavens, Dickson H. 1953. Letter to the Editor. En *Econometrica*. Vol. 21, no. 4, 630-632.

López Calafi, J.; A. Salvador y M. de la Guardia. 1985. Estudio bibliométrico de la literatura sobre la determinación de elementos metálicos en aceites lubricantes por espectroscopía atómica. En *Revista Española de Documentación Científica.* Vol. 8, no. 3, 201-213.

López-Muñoz, Francisco y G. Rubio Valladolid. 1995. La producción científica española en psiquiatría: un estudio bibliométrico de las publicaciones de circulación internacional durante el periodo 1980-1983. En *Anales de Psiquiatría.* Vol. 2, no. 2, 68-75.

López-Muñoz, Francisco; Jesús Boya; Fernando Marín y José Luis Calvo. 1996. Scientific research on the pineal gland and melatonin: a bibliometric study for the period 1966-1994. En *Journal of Pineal Research.* Vol. 20, no. 3, 115-124.

Lotka, Alfred J. 1926. The frequency distribution of scientific productivity. En *Journal of the Washington Academy of Sciences.* Vol. 16, no. 12, 317-323.

Loughner, William. 1992. Lotka's law and the Kolmogorov-Smirnov test: an error in calculation. En *Journal of the American Society for Information Science.* Vol. 43, no. 2, 149-150.

Nath, Ravinder and Wade M. Jackson. 1991. Productivity of management information systems researchers: does Lotka's law applied? En *Information Processing & Management*. Vol. 27, no. 2-3, 203-209.

Oliveira, Silas Masques de. 1983. Aplicação da lei de produtividade de autores de Lotka á literatura de Jaca. En *Revista de Biblioteconomia de Brasilia.* Vol. 11, no. 1, 125-130.

Oppenheim, C. 1986. Use of online databases in bibliometric studies. En International Online Information Meeting. (9th: 1985: London). Oxford: Learned Information. p. 355-364.

Pao, Miranda Lee. 1985. Lotka's law: a testing procedure. En *Information Processing & Management.* Vol. 21, no. 4, 305-320.

Pao, Miranda Lee. 1986. An empirical examination of Lotka's law. *Journal of the American Society for Information Science.* Vol. 37, no. 1, 26-33.

Potter, William Gray. 1981. Lotka's law revisited. En *Library Trends*. Vol. 30, no. 1, 21-39.

Urbizagástegui Alvarado, Rubén. 1999. La ley de Lotka y la literatura de Bibliometría. En *Investigación Bibliotecológica.* Vol. 13, no. 27, 125-141.

Urbizagástegui Alvarado, Rubén and Shelley Lane. 2004. Lotka's law: an annotated bibliography. Riverside, Calif., 2004. Unpublished.

Urbizagástegui Alvarado, Rubén y María Teresa Cortés. 2002. La productividad de autores en la *Revista Geológica de Chile*. En *Ciencia de la Información.* Vol. 33, no. 2, 24-36.

Vlachý, Jan. 1980. Evaluating the distribution of individual performance. En *Scientia Yugoslavica.* Vol. 6, no. 1-4, 267-275.

Williams, C. B. 1944. The number of publications written by biologists. En *Annals of Eugenics.* Vol. 12, 143-146.

Tabla 1: Valores críticos de la prueba Kolmogorov-Smirvnov

One-sided test	$p = 0.90$	0.95	0.975	0.99	0.995
Two-sided test	$p = 0.80$	0.90	0.95	0.98	0.99
$n = 1$	.900	.950	.975	.990	.995
2	.684	.776	.842	.900	.929
3	.565	.636	.708	.785	.829
4	.493	.565	.624	.689	.734
5	.447	.509	.563	.627	.669
6	.410	.468	.519	.577	.617
7	.381	.436	.483	.538	.576
8	.358	.410	.454	.507	.542
9	.339	.387	.430	.480	.513
10	.323	.369	.409	.457	.489
11	.308	.352	.391	.437	.468
12	.296	.338	.375	.419	.449
13	.285	.325	.361	.404	.432
14	.275	.314	.349	.390	.418
15	.266	.304	.338	.377	.404
16	.258	.295	.327	.366	.392
17	.250	.286	.318	.355	.381
18	.244	.279	.309	.346	.371
19	.237	.271	.301	.337	.361
20	.232	.265	.294	.329	.352
21	.226	.259	.287	.321	.344
22	.221	.253	.281	.314	.337
23	.216	.247	.275	.307	.330
24	.212	.242	.269	.301	.323
25	.208	.238	.264	.295	.317
26	.204	.233	.259	.290	.311
27	.200	.229	.254	.284	.305
28	.197	.225	.250	.279	.300
29	.193	.221	.246	.275	.295
30	.190	.218	.242	.270	.290
31	.187	.214	.238	.266	.285
32	.184	.211	.234	.262	.281
33	.182	.208	.231	.258	.277
34	.179	.205	.227	.254	.273
35	.177	.202	.224	.251	.269
36	.174	.199	.221	.247	.265
37	.172	.196	.218	.244	.262
38	.170	.194	.215	.241	.258
39	.168	.191	.213	.238	.255
40	.165	.189	.210	.235	.252
$n > 40$:	$\frac{1.07}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.52}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$